International Journal of Child-Computer Interaction 21 (2019) 37–53

Contents lists available at ScienceDirect

International Journal of Child-Computer Interaction

journal homepage: www.elsevier.com/locate/ijcci

Going beyond performance scores: Understanding cognitive–affective states in Kindergarteners and application of framework in classrooms

Priyashri K. Sridhar [a,b,*], Samantha W.T. Chan [a], Yvonne Chua [a], Yow Wei Quin [b], Suranga Nanayakkara [a]

[a] The University of Auckland, 70 Symonds Street, Auckland 1010, New Zealand
[b] Singapore University of Technology and Design, 8 Somapah Rd, Singapore 487372, Singapore

ARTICLE INFO

Article history:
Received 30 September 2018
Received in revised form 20 January 2019
Accepted 11 April 2019
Available online 16 April 2019

Keywords:
Emotion / affect
Children
Learning
Empirical study
Cognitive state

ABSTRACT

Cognitive–affective states during learning or interactions with technologies are related to the mental effort of the learner and / or the cognitive load imposed by the system. Despite the growing research on the importance of understanding cognitive–affective states and their relationship to learning, measurement of such states during the learning process is still unclear. While most assessments of learning and usability evaluations with Kindergartners and primary schoolers focus on performance, self-reports and inferring from observable behaviours, they provide limited insights into the cognitive load and emotional state during learning or interaction that are essential for a holistic picture of learning. Through a study with 18 Kindergartners, we explore the feasibility of understanding cognitive–affective states associated with mental effort by triangulating the data obtained from observations, physiological markers, self-reports and performance as they performed tasks of varying mental effort. We present findings on the reliable markers within these sources across tasks. Results reveal that such a triangulation offers deeper insights into the cognitive–affective state of the learner. As a follow-up study, we explored the feasibility of employing this method in classrooms with 36 children between 11–13 years to understand the effect of interactivity on learning across three viewing conditions namely, screen, magic window and immersive virtual reality (VR) using Google Cardboard. Results confirmed the feasibility of running such studies and the additional value of employing a host of measures to evaluate the learning. We believe this work would be a step towards better understanding of the learning process, thereby facilitating instruction that is more aligned with the learner's cognitive–affective architecture. Further, we believe that such methods have applicability in comprehensive usability / evaluation processes based on well-defined associations between child behaviour and child action.

1. Introduction

Educational psychologists have increasingly emphasized the importance of kindergarten education in a child's overall development. A significant way of creating enriching experiences comes from a thorough understanding of the cognitive–affective state of the child and following their learning behaviour. Extensive research on measuring cognitive load through self-reports [1] provide limited insight to the quantity of knowledge and no information on the learner's cognitive load or emotions during the learning process [2]. These measures when used alone are static (measured at a single point in time), thereby making them inappropriate for measuring variations in cognitive load over a continuous time frame. Furthermore, there are mixed views on the accuracy and reliability of self-reports especially with children [3].

In order to determine how to respond to the temporal and subtle changes of cognitive–affective states as well as improve the reliability of subjective responses, it is necessary to objectively measure the cognitive load of Kindergartners in real-time and in-situ. While physiological and neurological measures such as skin conductance, heart rate variability and Functional Near-Infrared Spectroscopy (fNIRS) have been explored in the context of cognitive load, such research is primarily focused on adults. To our knowledge there has been very limited exploration in understanding the physiological changes that correspond to cognitive–affective states during learning in Kindergartners. By determining the objective and subjective markers that correspond to increased cognitive load, we can help understand the learner's cognitive–affective state during learning.

As a first step in this direction, we investigated the feasibility of obtaining physiological measurements from Kindergartners.

* Corresponding author at: Singapore University of Technology and Design, 8 Somapah Rd, Singapore 487372, Singapore.
E-mail addresses: priya@ahlab.org (P.K. Sridhar), samantha@ahlab.org (S.W.T. Chan), yvonne@ahlab.org (Y. Chua), suranga@ahlab.org (S. Nanayakkara).

https://doi.org/10.1016/j.ijcci.2019.04.002
2212-8689/

We then conducted a controlled study with standardized cognitive tasks of varying difficulty to identify suitable physiological markers. We found that specific markers within skin conductance and heart rate are linked to increasing cognitive load. However, physiological measures are characterized by lack of specificity. This refers to the fact that while an increased in skin conductance responses may signal increased arousal or attention, it does not in anyway signal the valence of this emotion. An increased arousal maybe positive (as in happiness and excitement) or negative (as in frustration or confusion). Thus, it was challenging to infer the cognitive–affective state accompanying an increase in cognitive load, as experienced by the learner. Hence, we triangulated the physiological measures with observational data to understand the events and emotions that accompanied or triggered the physiological change. This triangulation helped us tease apart the pattern of physiological measures and revealed better insights into the cognitive–affective state. We further employed and tested the feasibility of this framework in a classroom to understand learning in children, by comparing interactive and passive modes using screen, Magic Window and immersive virtual reality. The contributions of this paper are four-fold:

1. Exploring the feasibility of obtaining physiological measurements of cognitive load from Kindergartners as they engage in cognitive tasks
2. Conducting a user study to obtain physiological, observational and performance measures, and triangulating the results from the three sources to better understand cognitive–affective states during performance of cognitive tasks.
3. Discussing insights on formulating and implementing triangulated study designs with Kindergartners.
4. Exploring feasibility and added value of employing physiological, behavioural and observational data in classrooms through an investigation of learning across interactive and passive modes.

2. Related work

2.1. Assessment of cognitive states in learning

One of the popular approaches of assessment of cognitive states has focused on cognitive load imposed by the learning or the mental effort exerted by the student and the demonstration of proficiency in a subject matter. York et al. [4] attempt to evaluate programs that work towards attaining such focus such as "Knowledge of Individual Students' Skills" (KISS) [5] that are based on how well teachers' ratings align with students' actual proficiencies. Teachers may also rely on observable behaviours in the classroom to infer about a child's learning status. While these offer insights into a learner's knowledge acquisition, they usually are conducted after lesson delivery and therefore do not offer much on what really happens during the learning process itself. Further observations are not always representative of a learner's skill/understanding (as the observation made at certain points only and may miss other behaviour) and are prone to be affected by observer prejudices and biases. As a result, there is not much information on how much mental effort is being exerted by a student on a learning task, whether the task imposes extreme cognitive load, the nature of the cognitive load as well as what are the range of emotions the student goes through. According to Kirschner [6], access to mental effort by learner enables timely intervention of the teacher to redesign instruction in a way that makes learning enjoyable to the child.

Other methods have been used to infer about cognitive load as well as how a student feels about a task. The subjective or self-reporting method [1] has been the most commonly used method with adults due to its convenience. However, for children below age 11, self-reports have low validity due to their limited language ability, reading age, motor skills, temperamental effects such as confidence, self-belief and desire to please [3]. Another important aspect of these questionnaires is the time of administration [7]. Most of the studies present the questionnaires after the learning has occurred [8,9]. As a result, there is a high possibility that the participant may provide an average estimate for the whole task that is affected by memory effects. This loses its purpose of capturing the dynamic and fluctuating nature of load that is imposed during learning [10].

The second common method of measuring cognitive load is using dual or secondary tasks [11,12] that draw from psychology where a secondary task is introduced along with the primary task of learning. However, as Yuksel et al. [10] point out, a major disadvantage of these tasks is their interference with the primary task especially when the primary task itself is complex and draws much of the learner's cognitive capacity [13].

With the inclusion of technology into the learning environments, there have been some encouraging explorations on human thinking and information processing abilities [2]. Many of the studies of cognitive load and learning outcome measures administer each of these measures either before or after test performance [14]. Even though they are static and considered unreliable [15], they continue to be popular in real-world contexts partly because single measures are easy to administer whereas other objective measures of cognitive load may require expensive and hard-to-use instrumentation. In contrast, Mayer et al. [16] have highlighted the need for direct measure of cognitive overload. Physiological measures such as skin conductance and heart-rate variability offer a direct measure of cognitive load [2,17].

2.1.1. Galvanic skin response (GSR)

Research on skin conductance looks at the skin conductance response (SCR) that is triggered by the action of sweat glands to an external stimulus. Researchers [18,19] have used GSR to differentiate between stress state and cognitive load state, and found correlations between the GSR signal and cognitive load. It has been shown that parameters of GSR are subject to user movements and the rise/fall is relative to the trigger event [20] while others find a weak relationship between skin conductance and cognitive load [21]. Ferreira et al. [22] used perceptual speed and visuo-spatial cognitive processing tasks and collected psychophysiological data in young and old adults that included GSR. With pre-schoolers, researchers have explored GSR as objective indicators of anxiety [23] or aggression [24]. Such work shows GSR as a potential physiological marker for different behaviours.

2.1.2. Heart rate variability

Cognitive load has been shown to have an effect on various components of Heart Rate Variability (HRV) such as mean heart rate (HR), breathing rate, low frequency (LF) and high frequency (HF) components of HRV [25–27]. People under high mental workload have reduced HF components [26]. The HF component of HRV is indicative of the parasympathetic influence on the heart and is high during rest. During high-attention tasks, absolute measures of LF and HF HRV power have been observed to decrease when compared to a baseline [27]. Mc Duff et al. [28] used remote HRV measures to monitor effect of cognitive workload on HRV and identified the LF and HF components of HRV to be the most indicative of cognitive stress.

Although physiological measures provide a direct measure of cognitive load, there are limitations when they are used by their own. Changes in GSR and HRV can also be mapped to other phenomenon such as changes in emotional states [29]. Bearing this in mind, we use the physiological measures in conjunction with other observational data and the performance across trials and over time.

2.2. Role of affect in identifying cognitive load

Even though affect has been proposed to play a great role in learning [30,31], it has not been explored as an extension of cognitive theory [2]. There has been consensus on the role of intrinsic vs. extrinsic influences, the influence of past pleasurable experiences and motivation on learning [32–35]. Some researchers have integrated both affective and cognitive components into motivation theories [36–38]. While these have provided insight into emotions and reaffirmed their role in learning, there is not much consensus on the kind of emotions involved in learning. Csikszentmihalyi [39] has emphasized the tendency for a pleasurable state of "flow" to accompany problem solving. Kort [40] attempted to list the emotions involved in learning and proposed a four quadrant model relating phases of learning in emotions. There have also been some scattered attempts by other researchers [41,42] to identify emotions in learning.

One of the reasons why affect is not explored much is that it is hard to measure. While it is easy to get performance scores and test the ability to transfer learning, it is harder to measure how the learner feels during the entire process [2]. Typically affect has been measured through questionnaires that ask how much pleasure, frustration or interest one felt [43–45]. Moreover, there are specialized instruments for evaluating the motivational characteristics of an instructor's classroom delivery [46] but as shared earlier, self-reports for children are not reliable and valid. Much of earlier work has focused on emotions from exaggerated expressions, making it hard to generalize them to typical learning situations. Kapoor et al. [47] attempted to build a system that detects surface level affect behaviour such as posture, eye-gaze, facial expressions and head movements using pressure sensors and gaze-tracking.

There is a strong need to understand affective states with respect to cognitive load in order to understand the learning state as well as suitably intervene or design interfaces that intervene appropriately. For example, a learner who makes mistakes while appearing engaged and curious is different from a learner who makes mistakes while fidgeting, frowning or displaying other anxious behaviours. While the former is important as it encourages exploration, the latter case demands some intervention, re-instructions and maybe feedback to facilitate learning. While there seems to be a general consensus among educators that interest and engagement are important factors in the learning process, much of this consensus is based on intuition and there is a need for studies that look at affect in learning and in relation to cognitive load.

2.3. Multimodal sensing in understanding cognitive–affective states in learning

There have several encouraging attempts that combine physiological, postural, voice, facial expressions, eye gaze, neurological measures to sense affect. According to Calvo and D'Mello [48], the value of each modality depends on the validity of the signal as a natural way to identify an affective state, the reliability of the signals in real-world environments, the time resolutions of the signal as it relates to the specific needs of the application and the cost and intrusiveness for the user. The underlying tenet of such multi-modal approaches is the belief that emotion is an embodiment that is best captured through multiple physiological and behavioural approaches. For example, anger is expected to be manifested via particular facial, vocal, and bodily expressions, changes in physiology such as increased heart rate, and is accompanied by instrumental action [49]. Therefore, combining different responses in time will provide a better estimate of the emotion that may have ensued [50]. Some of the researchers have employed a combination of methods to sense affect [48,51–53]. There are several challenges that come with any form of unisensory detection as discussed in each of the cases above and they only get amplified in multi-sensory approaches. Collecting multi-modal information in natural settings is rid with several noise and confounding variables. Nevertheless, the advantage of multi-modal human–computer interaction systems has been recognized and seen as the next step for the mostly uni-modal approaches currently used [54,55]. The reciprocal relationship between knowledge/goals of the learning and the emotions has been previously established [56]. Different affective states have been shown to be associated with cognitive processes with positive affect accompanying flexibility, creative thinking, efficient decision making and negative affect linked to narrower localized attention [57,58]. While theories on emotion have assumed the link with cognitive processes [59–62], they only offer general relationships without explaining the exact nature of emotions that accompany complex learning tasks. However, there is encouraging emergent research on more nuanced and systematic relationships between affective and cognitive states during complex learning with a focus on a wider set of emotions instead of just anxiety/motivation-based traits [48,63–66]. While some researchers focus on a broad set of emotions in learning, there have been deeper analysis on sub-sets of emotions especially during learning over shorter spans of time from 30–90 min [67–69]. In a study with the Auto-Tutor [53], learning gains were shown to be positively correlated with confusion and flow, but negatively correlated with boredom. Confusion was shown to play an important role in the learning process. The emotions that appear to be prominent in these learning sessions include boredom, engagement/flow, confusion, frustration, anxiety, curiosity, delight, and surprise [49]. While Kort's quadrants [40] link affect to learning events, the model on affective dynamics by D'Mello et al. [51], is theoretically grounded in perspectives highlighting the importance of goal appraisal, cognitive disequilibrium, and impasse resolution during learning and problem solving. Through two studies, they showed that learning is an emotionally charged experience with experiences of engagement/flow, confusion, boredom, and frustration. The authors found that the transitions between these states were systematic and not random. Even though an exact model of such affect or a corpus of such states does not exist, the role of these cognitive–affective states remains largely undisputed. Most of such work is set in computer learning environments or with intelligent tutors involving tasks such as problem-solving, reading comprehension and essay writing. Therefore, most work focuses on older students in high schools and universities. However, such attempts have encouraged explorations of affect and its underlying and overt mechanisms, and serve as inspiration for much of the research. It is precisely these efforts into building affective systems, based on a phenomenon that is poorly understood, that will aid in understanding of that very phenomenon [70].

2.4. Cognitive–affective states in usability evaluations in children

There has been much interest in measuring the usability and engagement of a novel interface that was being designed for children. For a long time now, including children in the design of new technologies, either as informants or design partners has been highlighted as being beneficial to understand users, gather design ideas and to test out new concepts [71,72]. Since recruiting children in testing of technology designed for them involves ethical concerns as well as methodological considerations, choosing the appropriate usability evaluation method (UEM) is important. UEMs that elicit only verbalization maybe too strict for Kindergartners thereby necessitating a need for some flexibility

in the approach to allow the child to express emotions, thoughts and opinions in activities. In addition, survey methods that adopt a question–answer process often are impacted by developmental effects; language, reading age, and motor abilities, as well as temperamental effects including confidence, self-belief and desire to please [3]. Also, when assessing children between 2–6 years for appeal or engagement, testers will need to closely observe behaviours such as sighing, smiling, or sliding under the table. Given the challenges with eliciting the exact response of how children truly feel, an objective method to measure cognitive–affective states as children go through the interaction with a novel interface or UEM maybe potentially useful. Therefore, in this study, we also explore the feasibility of obtaining physiological, behavioural and observational measures with Kindergartners keeping in mind its potential in conducting UEMs with children in the future.

3. Pilot study: Exploring feasibility and value of using physiological, behavioural and observation measures to understand cognitive–affective states during learning

The pilot was conducted to check the feasibility of the study design and collecting physiological parameters from Kindergartners. We selected two established sensors to obtain physiological measurements: Empatica E4 wristband [73] with dry electrodes and Consensys Shimmer3 GSR Kit[1] with pre-gelled electrodes.

With our initial exploration with some Kindergartners, we noticed that since Shimmer3 GSR had to be secured to the fingers, it limited the range of movement on the hand where the skin conductance was measured. In addition, participants tend to get distracted with the wires and the electrode placement especially when the tasks demanded key press. Hence, we used only the Empatica E4 for its ease of use, adjustability to children's wrist, and possibility of obtaining multiple measures simultaneously. We also custom built a mobile app which synced to the E4 band to visualize data in real-time and save them into a local database.

3.1. Method

3.1.1. Participants

Three English–Mandarin bilingual Kindergarteners (2 female, 1 male) participated in this study (M_{age} = 5.58, SD = 0.49, age range: 4 to 7 years). The children were recruited from a Kindergarten in a middle-class neighbourhood. The average level of parental highest education was a university degree.

3.1.2. Stimuli

We used standardized tasks that map onto retrieval of stored knowledge from long-term memory as well as executive functions namely, inhibition, flexibility and working memory. These skills have been shown to impact learning and elicit mental effort on part of the participant. These tasks elicited different levels of mental effort.

Johnson Woodcock IV Test: Three sections from the Test of Cognitive Abilities that constitute the Brief Intellectual Ability were used. The sub-tests included: verbal ability (antonyms and synonyms), verbal attention and number series. Verbal ability requires recall/retrieval, verbal attention is a test of working memory while the number series tests working memory as well as inhibition. The items in each sub-test are arranged in the increasing order of difficulty. In the sub-test on antonyms and synonyms, the child was asked to say the antonym and synonym of the stimulus item respectively. In the verbal attention task, the child was asked to repeat the animal and number combinations in the same order as presented by the experimenter. In the number series task, the child was asked to identify the missing number by understanding the pattern/sequence of the stimulus item. For every stimulus item, the child was given a maximum of 1 min to respond, failing which, the next item was presented. Each response was scored and each of the sub-test was terminated when the participant responded inaccurately or did not respond for six consecutive test items.

Executive Function (EF) Tasks: A computerized version of the Simon Task [74] and the Dimensional Change Card Sort (DCCS) test [75] were presented using E-Prime software [76]. The Simon tasks involves executive function skills of inhibition and to a small extent, working memory. In Simon Task, the subjects were presented with a red or a blue square on the screen. They were instructed to press a button on the corresponding side of the stimulus. There were three types of trials: congruent, incongruent and mixed. In the congruent trial (lowest load on inhibition), the presentation of the stimulus matched the side of the response key. In the incongruent trial (higher load on inhibition in the first half but reduces over time), the stimulus presentation was located on the side opposite to the response key. In the mixed trial (highest load on inhibition because the trials are all randomly mixed making prediction impossible), there were congruent and incongruent blocks in a random order. The participant was instructed to press the right key as quickly as possible. The DCCS task is a test of inhibition and flexibility. In DCCS, the children were required to sort through a series of bivalent test pictures first according to one dimension (colour) or another dimension (shape). There were two blocks (congruent and mixed) of 20 trials each. In the congruent block (lower load on inhibition and flexibility as the participant sorts according to the same dimension throughout), the participants sorted the stimuli according to colour only. In the mixed block (higher load on inhibition and flexibility as they can be asked to sort on colour and shape interchangeably and randomly), the two dimensions of colour and shape were used interchangeably. The participants were instructed to press one of the two keys to denote their response as quickly as possible. In both tasks, the reaction time and accuracy were calculated.

3.1.3. Procedure

The study was conducted in a quiet room in the participants' school to ensure familiarity of surroundings. The experimenter built rapport with every participant by participating in their classroom activities during play and art lessons. Following this phase, the participants were recruited for the study. The experimenter conducted one-on-one sessions where each participant wore the E4 wristband prior to the sessions. They first completed a baseline period of sitting quietly and relaxing for 3 min. Following this, they were asked to press some random keys on the keyboard repeatedly to check if the movement affected the measurements. After confirming that the key press did not affect the readings, they completed the Johnson Woodcock Tests and the EF tasks (both were counter-balanced). The tasks were completed over a span of two sessions each lasting around 30 min. Other ambient conditions such as room temperature and lighting were controlled across participants and sessions.

3.2. Findings

Feasibility of physiological measurement: The pilot study revealed that it is indeed possible to collect physiological data from Kindergartners as they perform tasks with varying cognitive load. E4 wristband was convenient to collect physiological data as the sensors are wireless and the same wristband offers heart rate sensing as well as skin conductance sensing. However, in spite of

[1] http://www.shimmersensing.com/products/shimmer3-wireless-gsr-sensor.

the adjustable strap and getting the tightest fit, the PPG sensor for heart rate measurement sometimes did not achieve good contact with the wrist of the participant owing to their small wrist size, which resulted in data loss. In order to avert this, we used a small pad of cloth near the strap to enable better contact of the PPG sensor on wrist.

Selection of physiological parameters: Based on a preliminary analyses, we found that body temperature did not change much across tasks as compared to baseline. Therefore, we dropped this measure for the main experimental study. The GSR measures, particularly, the number of Skin Conductance Responses (SCRs) and average amplitude of SCRs showed variations across baseline and tasks. In HRV measures, we found a change in mean heart rate (HR), low frequency power of HRV and high frequency power of HRV in tasks as compared to baseline. We therefore decided to use both the GSR and the HRV measures for the main experimental study.

Importance of baseline and consistent ambient conditions: We realized the need for baseline in similar ambient conditions and between the tasks for every participant. We noticed that at the end of one task, the measures in GSR and HRV shift from a resting baseline. Therefore, there is a need to take a break and bring the values back to resting baseline before proceeding to the next tasks in order to obtain a true measure of the change in physiology for every task.

Quality control of collected physiological data: The data collected through E4 (the SCRs and HRV) was compared with the time-matched video recording to rule out any responses resulting from large movements. This often showed up as unusually large peaks in SCRs that matched with such movements. For HRV data however, noise usually resulted in data loss accompanying large movements of hand or body but much of the data collected during non-noisy periods was of good quality. Therefore, triangulating with video was not only a way of identifying underlying cognitive–affective states but also a means of validating the physiological data to a reasonable extent.

4. Main study: Triangulating performance, observational and physiological measurements

Incorporating the findings from the pilot study, the objective of this study was to triangulate performance measures, observational data and physiological measurements to explore whether behavioural analysis and physiological data can indeed reveal more insights into the cognitive–affective state of the participant beyond just performance scores/accuracy.

4.1. Method

4.1.1. Participants, stimuli and procedure

Fifteen English–Mandarin bilingual preschoolers participated in this phase of the study (M_{age} = 5.23, SD = 0.73, age range: 4–6 years; 9 males, 6 females). They were recruited from the same school. The stimuli used were the same as those used in the pilot study and the physiological measures (HRV and GSR) were measured with E4 wrist band. In addition, we collected performance data (response time and accuracy) and video-recorded the sessions for later analysis of emotions and behaviour.

4.1.2. Dependent variables

We compared the following dependent variables for all the participants across the baseline and experimental (task) conditions:

- *Performance measures*: This included the percentage correct responses for all tasks. For the EF tasks, we also calculated the response time.
- *Galvanic skin response*: This includes the number of Skin Conductance Responses (SCRs), average amplitude of SCRs and the cumulative amplitude of the SCRs for baseline and experimental conditions.
- *Heart rate variability*: This includes mean heart rate (HR), heart rate variability (HRV), low frequency (LF) component of HRV and the high frequency (HF) component of HRV.
- *Observable behaviour*: The coded behaviours included emotion (such as happiness, sadness, anger, disgust, fear, surprise, contempt and neutral [77]), response latency, vocalizations/comments, head movement, postural change, gazing/eye movement and other signs.

4.2. Results

4.2.1. Performance measures

(a) Johnson Woodcock IV Tests (JW)

For the four sub-tasks of the JW (see Fig. 1), correct responses were assigned a score of "1" (denoted in green) while incorrect responses were assigned a score of "0" (denoted in red) across tasks. The scores were recorded manually by the experimenter. Each sub-test was administered until the participant made six consecutive errors. The last column denotes the performance score (number of correct responses and number of stimuli presented). If one were to merely look at the performance scores and performance accuracy (Fig. 3), they do not reveal much about the performance pattern. While some participants had a series of all correct responses followed by six consecutive incorrect responses (e.g., P12 for the Number Series sub-task in Fig. 1a), others had some incorrect responses right at the start that was then followed by correct responses (e.g., P11 for the Synonyms sub-task in Fig. 1c). The performance plot reveals a better picture of where the errors are.

(b) Executive Function (EF) Tasks

Similar to the JW tasks, the EF tasks were also scored "0" for incorrect and "1" for correct responses. In addition to the accuracy, the time taken to complete each sub-task was calculated. Fig. 2 reveals the response pattern for each participant across trials. It was noticed that most children performed all the Simon Tasks with at least 95% accuracy (Fig. 3). As is the case with JW tasks, the total correct score reveals nothing about where the participants made errors. As expected, the participants took a longer time to complete the DCCS mixed block which requires them to exert cognitive flexibility compared to the DCCS Congruent task (see Fig. 4). The performance scores and accuracy in percentage (Fig. 3) is aligned with the time taken to complete the task (Fig. 4). However, as the difficulty is not increasing in order like the JW tasks, one cannot see a pattern here. While performance scores offer an overall picture of how "well" a participant performed, it does not tell us much about what the participant experienced as they went through the tasks, the pain points and their emotions as they faced easy compared to difficult test items.

4.2.2. Galvanic skin response measures

We analysed three measures of skin conductance across the tasks: (a) the number of Skin Conductance Responses (SCRs) (Fig. 5), (b) the cumulative amplitude of SCRs (Fig. 6) and (c) average amplitude of the SCRs. Any change in skin conductance greater than 0.01 microsiemens (μS) was considered as an SCR [78]. A continuous decomposition analysis was performed and normalized scores were computed from Ledalab toolkit for Matlab 2016b [79]. Paired t-tests were conducted to compare these measures across the tasks and baseline.

(a) Number of SCRs

There was data loss from one participant due to issues with the sensor contact and we excluded that participant from GSR

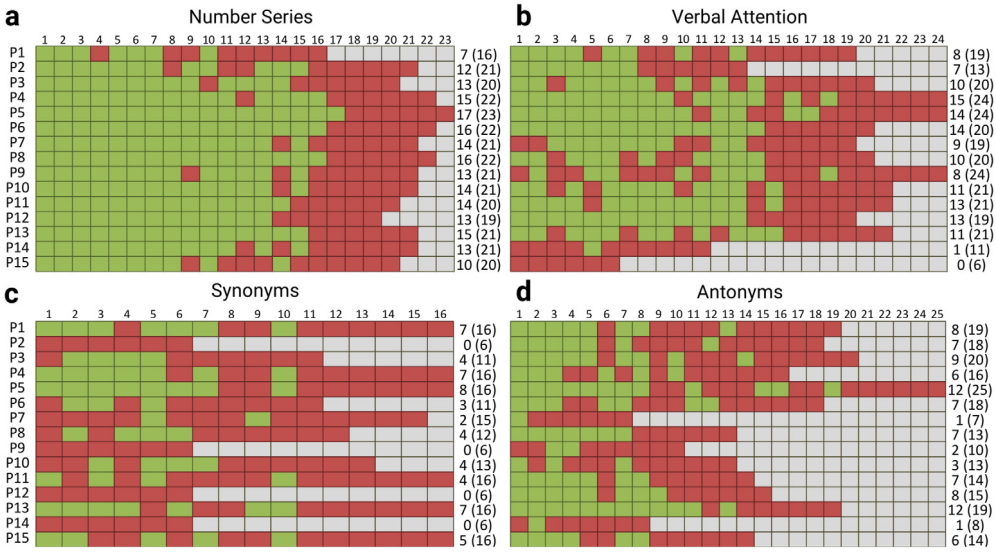

Fig. 1. Graphs of performance pattern across the four sub-tasks in the Johnson Woodcock Battery: (a) Number Series, (b) Verbal Attention, (c) Synonyms and (d) Antonyms. The green boxes denote correct response, red boxes denote incorrect responses. The grey boxes denote the questions that were not presented. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

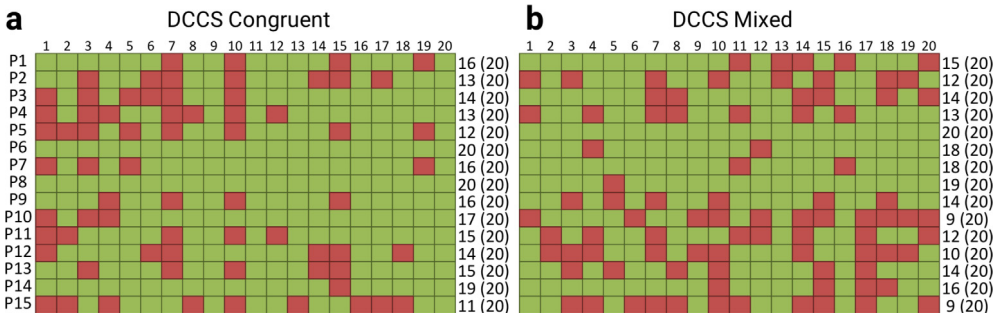

Fig. 2. Graphs of performance pattern across the EF tasks: (a) DCCS Congruent and (b) DCCS Mixed. The green boxes denote correct response and red boxes denote incorrect response. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

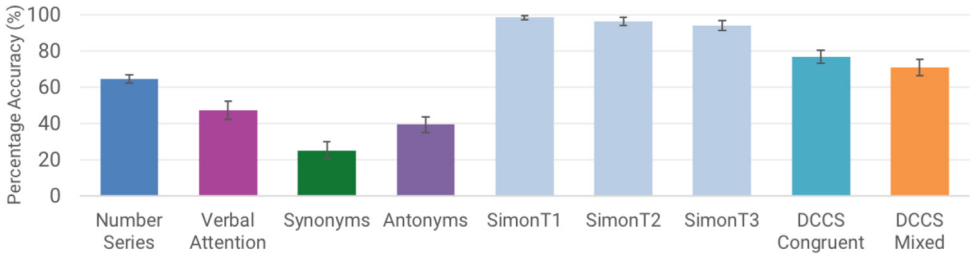

Fig. 3. Graph showing the mean accuracy in percentage + standard error (SE) across tasks.

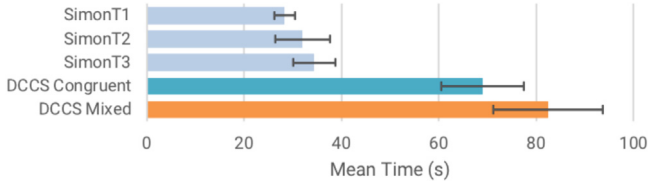

Fig. 4. Graph showing the total mean time taken + SE to complete the EF tasks.

analysis. Among the Johnson Woodcock sub-tasks, the number of skin conductance responses (Fig. 5) was significantly higher than baseline for synonyms ($t(13) = -2.608$, $p = .022$), verbal attention ($t(13) = -2.352$, $p = .018$) and number series ($t(13) = -3.97$, $p = .008$) sub-tasks of the Johnson Woodcock tests as compared to the baseline. Among the EF tasks, there were significantly more SCRs in DCCS-mixed block as compared to the baseline ($t(13) = -7.128$, $p = .00$). There was a marginally significant higher number of SCRs in the mixed Simon block than the baseline ($t(13) = -1.970$, $p = .061$). No such differences were

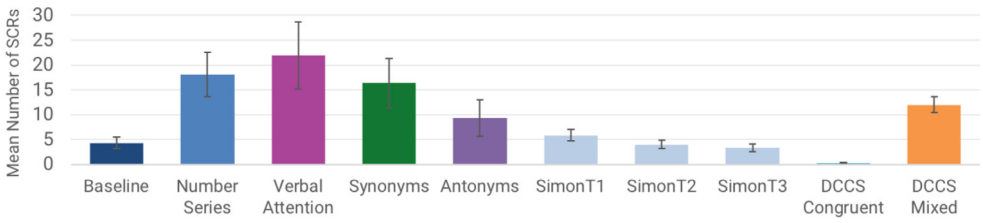

Fig. 5. Plot of mean number of SCRs + SE across tasks.

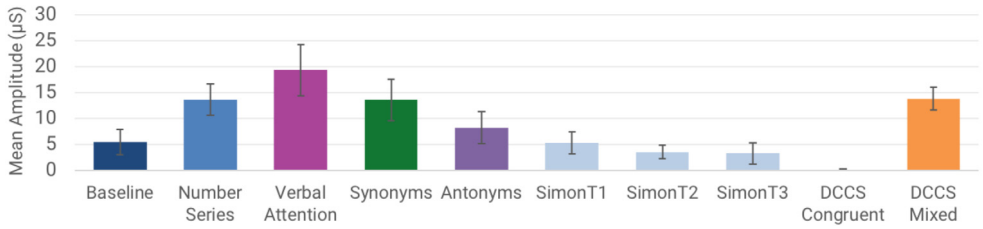

Fig. 6. Plot of mean cumulative amplitude of SCRs + SE (in μS) across tasks.

observed for the DCCS Congruent block as well as the congruent and incongruent trials of Simon Task. The DCCS Congruent block has relatively less cognitive load as compared to the mixed block that may have resulted in no significant difference between the groups. The skin conductance may not have been sensitive to the demands placed by the Simon task as the blocks were of very short duration and easier as compared to the mixed block for which differences were found. When compared to the performance measures, it can be noted that when a task is easy, the number of SCRs is lower. This is best illustrated with the DCCS tasks where the number of SCRs for mixed block (challenging) is higher than the congruent (easier) block.

(b) Cumulative SCR amplitude

The cumulative GSR amplitude refers to the sum of all the SCRs. There was a significant increase in the cumulative SCR amplitude compared to the baseline in the mixed DCCS block ($t(13) = 2.199$, $p = .001$); verbal attention task ($t(13) = -2.123$, $p = .0524$) and the Number Series task ($t(13) = -1.842$, $p = .056$). Even though synonyms had a significantly higher number of SCRs than the baseline, they were probably not of very high amplitude. Although graphs of SCR amplitude (Fig. 6) emphasizes the magnitude of difference, the emotional response to the demands of the task is unknown.

(c) Average amplitude of SCRs

We did not find any significant difference in the average value of SCRs across tasks.

GSR analysis reveals whether a certain marker is sensitive to cognitive load in Kindergartners and to some extent the amount of cognitive load imposed by different tasks that tap on different cognitive resources. Since GSR can be mapped to variety of emotional states such as excitement, frustration and engagement, there is a need to supplement this with behavioural observations to get a complete picture of cognitive–affective state.

4.2.3. Heart rate measures

We analysed various measures using Kubios HRV 2.2 [80] on Matlab 2016b.[2] The analysed measures included: mean heart rate (HR), mean inter-beat-intervals (RR), heart rate variability (HRV) and low frequency (LF) and high frequency (HF) components of heart rate variability. Overall, we found that the low frequency component of HRV as shown in Fig. 7 (modulated by sympathetic and parasympathetic activity) was significantly higher than baseline values for synonyms ($t(14) = -2.361$, $p = .015$), antonyms ($t(14) = 2.437$, $p = .0168$), and DCCS Mixed task ($t(14) = -3.378$, $p = .005$) and marginally significant for number series ($t(14) = 1.922$, $p = .054$). We did not find any significant changes in the parasympathetic activity measured through HF component (Fig. 8). This may be owing to the age group of the participants as they may not truly experience relaxation when on a task. While our analysis shows that HRV measures maybe sensitive to cognitive load in Kindergartners, they alone do not offer complete picture of how this load was perceived and whether the load resulted in frustration or encouraged them to be more curious and explore.

4.2.4. Behavioural video analysis

In addition to the manual coding, video recordings of the sessions were analysed using an application made through the Microsoft Emotion API,[3] which uses facial expressions detected from image frames of the video as input into Microsoft's cloud-based machine learning algorithm for emotion estimation. It outputs confidence scores (normalized to sum up to one) across 8 emotions (happiness, anger, sadness, fear, contempt, surprise, disgust and neutral).

Therefore while the toolkit does not specify the exact features, the facial expressions are used as a proxy for emotions. We wanted to explore the possibility of deriving emotions non-manually and integrating it with the rest of the data from the physiological sensors. However, the Microsoft API has not been previously validated on young children. The video recordings were independently coded by two researchers who have experience working with children. The coders inferred affect/emotions based on observable behaviours such as facial expressions. The coders therefore coded every performance trial for facial expression, head movements, body language, comments made and other overt behaviours. An affective coding index was not followed, however, the coders were asked to assign an affect/emotion based on their observation of facial expressions and other non-verbal behaviour. After coding, the two coders discussed their responses to arrive at one common agreed upon affect. The coded behaviour was first organized into those for correct responses and incorrect responses. This was done to compare differences in the observed affect when the responses were correct vs. incorrect. The differences are outline in Table 1. These were then

[2] https://www.kubios.com.

[3] https://azure.microsoft.com/en-us/services/cognitive-services/emotion/.

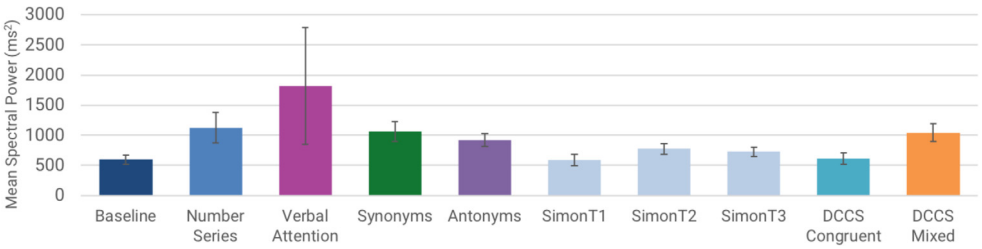

Fig. 7. Plot of mean Low Frequency power + SE (in ms^2) across tasks.

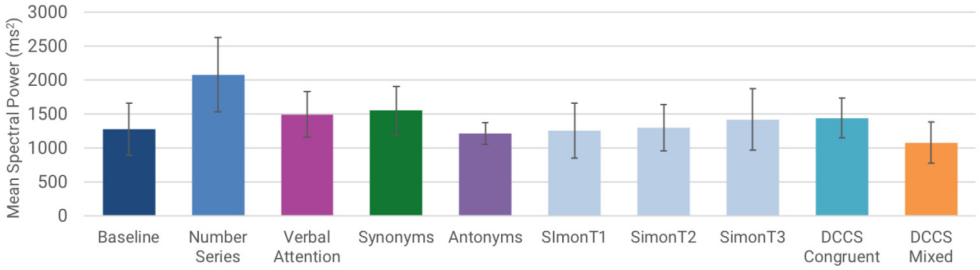

Fig. 8. Plot of mean High Frequency power + SE (in ms^2) across tasks.

Table 1
Manual video coding of behaviour.

Feature	Correct response	Incorrect response
Facial expressions/Assigned affect	Happy, calm, neutral, confident	Sad, frustrated, irritated, bored, confused, anxious, curious
Response latency	Fast, occasional pauses	Filled pauses "Oh no", "How much more?", "Can we play another game?", "When can I go?"
Head movements	Gentle leaning in	Head tilt towards floor lay head down on table looking away repeated head shakes
Postural change	Straight and alert sometimes casual	Rigidity move to edge of seat standing up leaning all the way back pressing palms against table
Gazing	Looking towards experimenter for affirmation, good eye contact	Looking to experimenter for affirmation, gazing away, looking elsewhere as an attempt to disengage from stress
Response time	Usually fast except when child tried justify an answer	Slow and laboured sometimes

further categorized into facial expressions and assigned affect, response latency (time taken to respond after a stimulus was presented), head movements, utterances, eye gazing and other overt behavioural signs [81] (Table 1).

We illustrate three sample outputs from the Emotion API. Fig. 9a shows the breakdown of emotions over time for number series task from participant P8. As the task involves pattern recognition, there are a lot of surprise peaks as new stimuli are presented. Similarly, Fig. 9b shows the emotions for participant P1 for the DCCS Mixed block. Since the difficulty is mixed across trials, there are a lot of surprise peaks as she switches between conditions imposed by the task but there are a lot of happiness peaks as she gets her answer right and makes comments throughout the session. Similarly, participant P13 (Fig. 9c) appears quite calm overtly as she attempts the verbal attention task. The application using Microsoft Emotion API is able to detect micro-expressions of sadness that increases as the task difficulty increases.

The Emotion API and manual coding of emotions were in-sync with each other. While such behavioural analysis reveals insights into the course of emotions as the participants faced different tasks with different difficulties, they alone do not offer insights into whether the child displayed an emotion in the presence of cognitive load.

4.3. Putting them all together: Triangulating physiological, performance and behavioural measures

We observed that while every measure offers a different perspective to the mental effort, they do not offer an entire picture of the individual or the process when used alone. For example, finding a one-to-one mapping of GSR to an emotional/psychological response is impossible as GSR can be high for positive and negative valence emotions. However, when GSR is evaluated in conjunction with HRV, it narrows down the list of possible emotions. For example, orienting, startle and defensive responses both elicit a high GSR. However, only startle and defensive responses are accompanied by increased heart rate [82]. Hedman [83] calls for a thick psychophysiological approach to understand events in the world using quantitative measures, external influences that may cause a physiological response and internal influence that refers to the meaning of that measure. In his studies, he uses video recordings in conjunction with skin conductance responses. In our study, we triangulated the physiological measure and performance data with our observations to explore if such an approach would offer us more holistic insights to the cognitive–affective state of the children. In the following two paragraphs, we describe two exemplary cases.

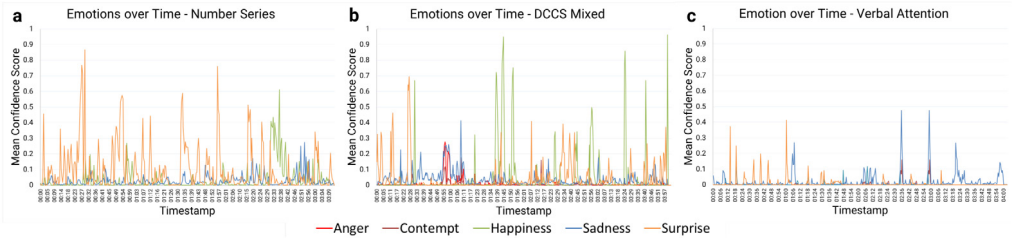

Fig. 9. Graphs of Emotion over Time for: (a) Number Series, (b) DCCS Mixed and (c) Verbal Attention. Analysed for three different participants.

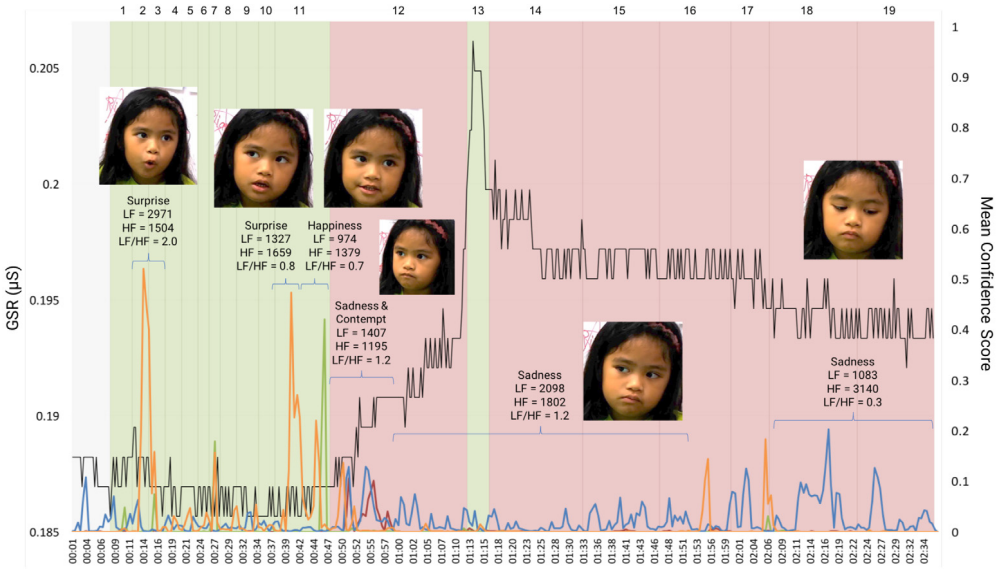

Fig. 10. Triangulation of measures (— representing GSR, other colours representing emotions and HRV values annotated. Orange line represents surprise, green line represents happiness and blue line represents sadness. The green and the red shaded regions represent correct and incorrect responses respectively.) for Participant 13 during the Antonyms task. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

4.3.1. Case 1

Fig. 10 shows the response of participant P13 as she went through the Antonyms sub-task of the JW. The antonyms sub-task taps onto the recall / retrieval of previously stored knowledge from long-term memory. Thus, the mental effort on part of the child is concerned with retrieval of pre-existing knowledge. The green shaded areas with the item numbers on top denote the correct responses while the area in red denoted the incorrect responses. We can see that neither all correct responses nor incorrect responses share the same physiological and emotional characteristics. In addition, P13 had no SCRs (every change in the skin conductance level was less than 0.01 microsiemens). The first high peak of surprise at around 15 s of the procedure is characterized by surprise accompanied by a high LF component of HRV, indicating mental load. This may have been due to processing instructions and the novelty of the task. The LF/HF ratio is also high, indicating a higher sympathetic activation and cognitive stress [28].

However, as she gets familiar with the task there is still an element of surprise around item 10–11 but we can notice that the mental effort/load imposed by the task has reduced as shown by the LF/HF ratio and reduced LF power. As she gets comfortable with the task and has a spate of correct response, she demonstrates happiness and an even lower LF power and LF/HF ratio. It is here that the emotions start changing towards sadness. When she faces a difficult question to which she is unsure of the answer, she demonstrates sadness and a very small proportion of contempt with LF of 1407, HF of 1195 and LF/HF of 1.17. Following this, there are more periods of sadness but they are characterized by different heart rate measures. For instance, the sadness at Item 15 has a high LF, indicating higher cognitive load, and an attempt to think and solve the question. Towards the flag end of the test, even though the emotion is still sadness, she seems to have given up which is also reflected in the reduced LF power as well as sympathetic activity that highlight that she is not stressed or even exerting much mental effort anymore. The presence of HRV measures compensated for the lack of GSR values. This emphasizes the need to triangulate the measures and understand them in light of the performance, the emotions and approach adopted by the child. This also finds affirmation with her body posture that seems to be more alert with a gentle leaning in and then becoming rigid as test items become more challenging. Towards the end, her posture becomes more relaxed as she realizes that she does not seem to know the answers to the questions anymore. By looking at this combined representation, we get a much clearer idea of what P13 went through during the procedure.

4.3.2. Case 2

Now consider participant P8 (Fig. 11) as he attempted the Number Series sub task of JW. The number series task tests the working memory and pattern recognition of the participant. As the task progresses, it requires the user to hold more information in their working memory as they process and derive the number that should follow the series presented. The green and the red shaded regions represent correct and incorrect responses respectively. Unlike P13, P8 demonstrates SCRs throughout. He starts off with a very overt expression of surprise characterized by a

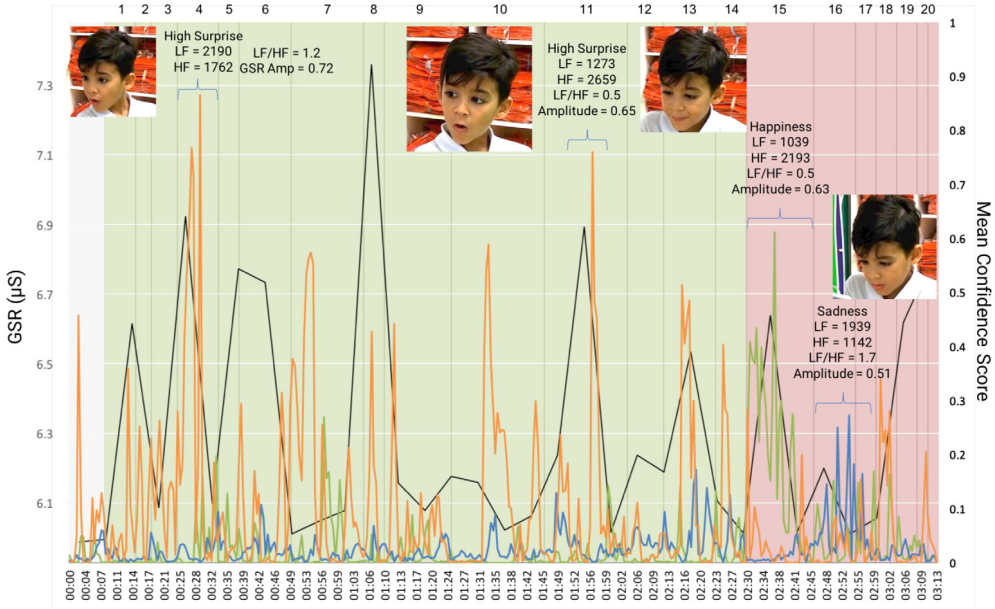

Fig. 11. Triangulation of measures for Participant 8 during the Number Series task.

high LF/HF ratio that again can be attributed to the excitement that accompanies the onset of a new task. This is corroborated by presence of a strong SCR which is also an indicator of arousal and excitement. Since no other emotion such as fear has a high value, one can ascertain that this surprise maybe more of excitement. Somewhere midway into the task, there is another surprise event that is also followed by an almost equal SCR like the previous one. However, the HRV LF and LF/HF measures which are an indicator of cognitive stress have now dropped lower. Once P8 reaches the incorrect response region, there is an event of "happiness" right at the start that may be attributed to having had a series of correct responses or being unaware of the first incorrect response he makes. This is to some extent corroborated by the HRV measures that still show very low cognitive stress on part of the child. The SCR remains almost the same. However, towards the end of the tasks after a series of incorrect/no responses, he seems less sure of his answers and the emotion of sadness is quite strong. That he is aware of his wrong answers and finds the test items difficult is shown by the increase in the LF and LF/HF measures of HRV. SCR is present as well albeit slightly lower in amplitude. Expression of emotions is highly subjective and also varies across emotions. While the expression of surprise in P8 is very obvious, P13's surprise is not as overt. But, P13's expression of sadness is more marked than that of P8. The emotion API is able to draw these emotions out to a good extent and they correspond well with the physiological measures.

No matter how efficiently an emotion is recognized, the information of mental effort is important to understand the child's approach according to the task difficulty. Therefore, tagging behavioural events during, and even before and after the physiological responses may facilitate a better understanding of the child's state. If a child approached a difficult problem with curiosity and happiness in spite of experiencing cognitive stress, then it is exploratory and needs to be encouraged. But if the child approaches a problem with sadness and shows a high cognitive stress, there may be a need for some feedback/intervention. Such insights are best attained by triangulating measures from different sources.

5. Exploring the feasibility and value of employing physiological, behavioural and observational measures in classrooms

Having ascertained the feasibility of obtaining physiological, behavioural and observational measures from children and the insights they offer towards understanding underlying cognitive–affective states across varying levels of mental effort, we wanted to test the feasibility of employing such a framework in a classroom in a group setting. Through this study, we wanted to understand the influence of immersion and interactivity in learning.

5.1. Method

5.1.1. Participants

Thirty-six Year 7 and 8 students (23 male, 13 female) from a low decile school in Auckland participated in the study. Students ranged from 11 to 13 years old (mean age = 11.97 years, SD = 0.71).

5.1.2. Learning content

The learning content used in the study was co-designed with two teachers from a blended Year 7 and Year 8 classroom in the primary school to supplement their inquiry (science) lessons. The teachers chose the topic to be bridges and structures with a focus on understanding how structures are designed, their purpose and what makes a good structure. Two types of learning content related to the topic were created: *Build the First Bridge (BB)* and *The Time Travelling Mailman (MM)*. Detailed descriptions can be found in Appendices A.1 and A.2.

5.1.3. Conditions

A between subjects design was used to investigate the effect of two levels of interactivity: *passive* consumption of video content and learner-paced *interactive* exploration. We investigated these two levels of interactivity in three viewing conditions providing different levels of immersion using devices that are readily available to classrooms today:

1. *Screen (SC):* Content displayed on a normal tablet screen (iPad).
2. *Magic Window (MW):* Content displayed in 'Magic Window' mode on the tablet where the tablet becomes a window into a virtual world and moving the tablet around physically in space allows the user to see different parts of the virtual world.
3. *Immersive Virtual Reality (IVR):* Content displayed on a mobile phone placed inside a mobile Virtual Reality viewer (Google Cardboard).

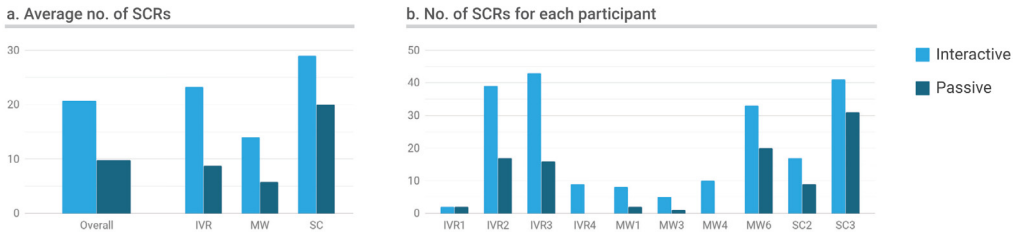

Fig. 12. Number of Skin Conductance Responses (SCRs) when viewing interactive and passive content: (a) averaged overall and by viewing condition and (b) reported individually.

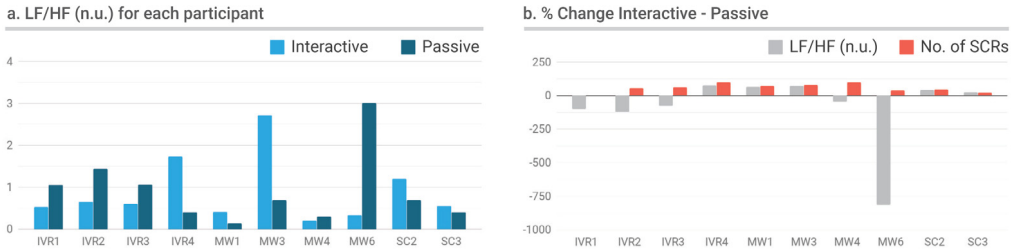

Fig. 13. (a) Average HRV LF/HF ratio of each participant when viewing interactive and passive content, (b) Percentage change of LF/HF ratio and no. of SCRs from passive to interactive content.

5.1.4. Dependent variables

We used the following measures:

- *Physiological measures*: The Empatica E4 wristband was used to collect electrodermal activity and heart rate data. Due to a limited number of devices, E4 data was collected from only a subset of participants (n = 22).
- *Observations*: Video and screen recordings were taken to aid in the triangulation of physiological measures.
- *Self-report*: The Smileyometer, Again–Again, and Fun Sorter (best to worst learning; most to least fun) components of the fun toolkit [84] were used.
- *Learning questionnaires*: Learning questionnaires were designed with teacher feedback, based on the specific learning goals for each content, as well as overall learning goals for the topic. The list of questions can be found in Appendix A.3.

5.1.5. Procedure

Students were separated into 2 groups for the study according to their existing inquiry class groupings (n = 12, n = 24). All students were present in the same room, clustered at different tables based on viewing condition. Six researchers were present on the day of the study and the sessions were video-taped from different points of view in the classroom.

Students first completed pre-test learning questionnaires, then viewed the two types of learning content, completing questions relevant to each type of content immediately after viewing. One week later, we returned to the class and students completed learning questionnaires again. Questionnaires were completed in pairs as that was the way students normally worked in the classroom.

Due to a limited number of devices, E4 data was only collected from a subset of students. These students wore the wristbands and completed a baseline period of sitting quietly and relaxing for 5 min before starting the pre-test questionnaires. They wore the E4s throughout the session.

5.2. Results and discussion

5.2.1. Physiological measures

We extracted some physiological markers that had shown to be potentially sensitive to engagement and increased mental effort in children in our first study with Kindergartners. For electrodermal activity, we elicited skin conductance responses (SCRs) that refer to peaks in skin conductance and correspond with increased arousal/excitement. From the heart rate measures, we extracted low frequency (LF) and high frequency (HF) component of heart rate variability that were calculated as the area under the Power Spectrum Density curve corresponding to 0.04–0.15 Hz and 0.15–0.4 Hz respectively. We normalized the LF and HF to minimize impact of the difference in total power and computed the LF/HF ratio that provides an estimate of sympathetic modulation and has been shown to be a proxy of mental effort and cognitive load [28]. A higher HRV LF/HF value corresponds to increased mental effort.

To understand the influence of interactivity on learning, we compared SCR data only for participants who had recorded data for both interactive and passive sessions (n = 10). We found that all participants showed higher number of SCRs as they interacted with interactive content (Fig. 12) indicating increased engagement as compared to the passive condition. In order to understand how this engagement elicited different levels of mental effort/ induced cognitive load, we used HRV LF/HF.

Results revealed some individual differences (Fig. 13a). In the IVR condition, 3 of 4 participants had a lower LF/HF ratio indicating lesser mental effort in the interactive condition while the MW condition was divided equally with 2 participants expending more mental effort in each of the conditions. Both participants in the screen condition had a higher LF/HF ratio in the interactive condition. By comparing this with the SCRs (Fig. 13b) we found that while interactivity may have elicited increased engagement across all conditions and participants, the mental effort applied seemed to vary across conditions.

To understand the influence of immersion on learning, we compared SCRs and HRV LF/HF between viewing conditions. We did not find any trends and there were large individual differences between participants (Fig. 14).

5.2.2. Performance measures

Having seen higher engagement for interactivity, we analysed their performance scores as a more direct measure of their learning outcome. Students viewing interactive content showed larger performance gains between delayed and pre-tests across

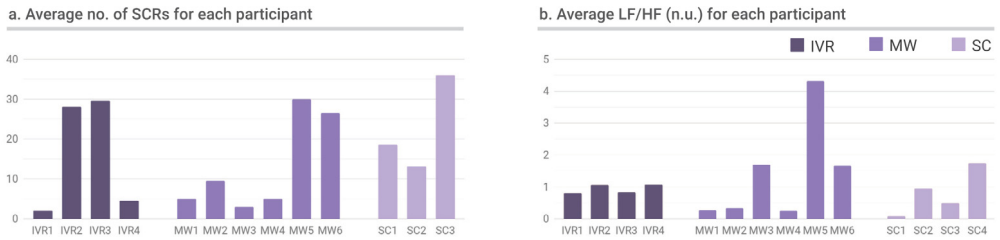

Fig. 14. (a) Number of SCRs and (b) Average HRV LF/HF ratio for participants in each viewing condition.

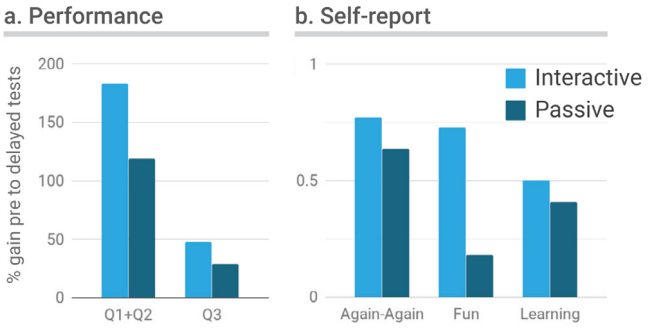

Fig. 15. (a) Percentage gain in performance scores from pre- to delayed tests, (b) Self-report scores from the Fun Toolkit: in Again–Again participants are asked if they would like to watch the video again (No(0), Maybe(0.5), Yes(1)); for the Fun Sorter participants arrange the 2 types of content they viewed in order of Least(0) to Most Fun(1) and Worst(0) to Best Learning(1).

content-specific questions: Q1 and 2 for BB content and Q3 for MM content (Fig. 15a). Overall, the interactive condition elicited better performance scores than the passive condition for both content types further reinforcing higher engagement with better scores irrespective of the mental effort.

We did not find a clear difference between the combined performance scores for each of the different viewing conditions for both pre–post as well as pre-delayed gains (Fig. 16a and b).

5.2.3. Self-reported measures

We finally compared the above measures with student self-reported scores to understand if students perceived better enjoyment with interactive and immersive content. We found that they rated the content higher on Again–Again and the Fun Sorter for the interactive condition (Fig. 15b).

Students in more immersive viewing conditions also rated the content higher on the Smileyometer and Again–Again (Fig. 16c).

5.2.4. Advantages of having multiple measures

HRV and Video data supplementing SCRs: There were large individual differences in SCRs and cognitive load (HRV LF/HF). Examining data on an individual level revealed insights into students' experience of the content. IVR1 and IVR4 both viewed the same type of interactive content in IVR. IVR1 had lesser engagement (2 SCRs) than IVR4 (9 SCRs) and a rise and fall of cognitive load over time while IVR4 had a continuous increase in effort over time. The screen recording for IVR1 revealed that higher cognitive load occurred during periods of multiple processing like listening to explanations, scanning choices and choosing a response. Interestingly the highest cognitive load mapped to the time when he made an error and was given corrective feedback and an explanation.

SCRs supplementing Video data: The obstruction of the child's eyes by the mobile viewer in the IVR condition posed difficulties in coding facial expressions. E4 data helped to supplement our understanding of covert behaviours, for example IVR2 despite appearing calm and reserved when viewing interactive content without displaying salient events in the video, had a high number of steadily increasing SCRs thereby telling us she was attentive/aroused.

Video data supplementing SCRs: Video data helped in understanding individual differences in SCRs within viewing conditions as it showed that students responded to the same viewing condition very differently. In the Magic Window condition there was a considerable difference in how much students moved around when viewing content, with some staying seated and relatively still (MW1, 4) and others moving around constantly in their seats (MW2, 3) or even getting up to walk (MW5, 6).

5.2.5. Feedback from teachers

We had some encouraging feedback from teachers on the implementation of such a framework in classrooms. The most commonly used method to assess student's states was through the use of self-reports, observations and listening to any issues/concerns they may have. One of the teachers also reported hat since she knows certain children need more guidance, she tends to pay more attention to them. However, they acknowledged that they cannot observe all students all the time and even though having a student's view is good as it gives them agency, it is not always reliable. One of the teachers also said that while she knows her students well, she sometimes has to second-guess what they are feeling. The teachers had several ideas on how this would help them in their classrooms. One teacher said that she would incorporate this as part of her formative assessments and use them during workshops and working with smaller groups and plan lesson plans accordingly. She said she would love to have

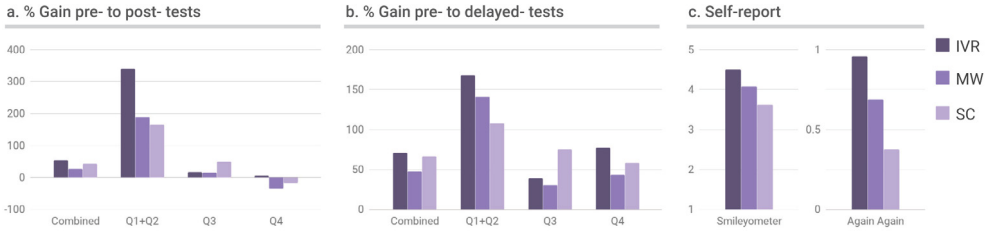

Fig. 16. (a) Percentage gain in performance scores from pre to post tests and (b) from pre to delayed tests, (c) Self report scores from the Fun Toolkit: in the Smileyometer participants rated the video on a scale of 1 (Awful) to 5 (Brilliant).

continuous visualization so she can even track students whom she cannot continuously observe. Another teacher felt encouraged hat the technology is wearable (except for cameras) and most students would not have issues wearing it since some of them are already wearing fitness trackers. She said she would use this information to match peers, change seats, add some story-telling or have visual charts for topics that are boring or challenging. She felt that such a framework would promote teacher agency (as the teacher would be able to drive a lot of the changes) and also student agency if she could get them to monitor their own states and also look out for their peers. One of the kindergarten teachers was concerned if the material of these wearables were safe for long term use by children. Another of the concerns a teacher had was the amount of time and effort she had to put in to interpret the data. All the three teachers suggested that visualizations that display information clearly (preferably on a big screen) would be really helpful.

5.2.6. Other considerations

Running controlled studies in classrooms: In spite of having had multiple researchers to help run the study, there were a number of confounding factors that arose from the dynamics of a classroom such as communication between students and other distractions. This required an in-depth pre-processing of physiological data with the video-tapes to rule out all responses that were associated with noise and in the process exclude participants.

Rethinking pre–post-performance evaluations: As student dissatisfaction with completing repeated questionnaires produced a confounding effect on post-test scores, we suggest designing performance evaluations in a more engaging manner, for example as a game or puzzle.

Cost of setup: The cost of deploying the E4 wristband with just a subset of students in a class was very expensive. Therefore, using them on a day to day basis with a large classroom is not financially feasible. However, as wearable technology gets more cost-efficient, we hope this will be alleviated to a large extent in the future.

6. Discussion

6.1. Specific challenges in eliciting Galvanic skin response data during learning from children

The pilot and main studies confirmed some of the previous guidelines when recording GSR from younger participants [85]. In addition, it also brought forth some adaptations to bear in mind when conducting such studies especially in longer sessions involving learning.

1. Restricted range of stimuli evoke SCR in children. Therefore, it was important to have different tasks and different kinds of trials in order to capture what GSR is more sensitive to.
2. Capturing non-specific SCR (NS-SCRs) especially for learning tasks is recommended. NS-SCRs are those responses that are not associated with discrete stimuli. Such responses are not measured in terms of amplitude but rather as number of SCRs per minute or over the time period of activity. A minimum value must be specified as the threshold. Current sensors allow for setting thresholds as low as 0.01 micro-Siemens that are more appropriate for small children as the NS-SCRs with no sudden external stimulus do not evoke the same intensity of response as in the case of sudden stimuli.
3. Monitoring any drastic changes in the tonic component of SCR is recommended. According to Dawson et al. [86], increases in NS-SCRs are attributed to (a) increase in tonic arousal, energy regulation or mobilization, (b) attentional and information processing, or (c) stress and affect. The tonic arousal can be ruled out by looking at any abrupt increases in skin conductance levels (SCLs) over time windows.
4. Need wireless real-time connectivity to offer real-time interventions: Since the long term vision of such an endeavour is to be able to deploy such solutions in the classrooms in the following years, we wanted to ensure we used methods that allowed for continuous real-time capturing of data.
5. SCRs in general are reported to appear later in children and even though they develop fairly well by 5–6 years [85] as we noticed in the responses in the study, it is recommended to include video taping of the sessions in order to capture other signs when physiological signals are missing. Further, since the SCR does not offer data on valence, the video-taping will facilitate interpretation of SCRs to some extent.
6. We observed that motor activity does not affect SCRs as evidenced when children were asked to perform random key press/ exhibited any other movements. In addition, it has been shown that normal motor movements do not evoke SCRs unless the person really jerks due to an emotion/ external stimulus like loud noise, there is no evidence showing that movement affects this [85]. Even if the SCRs are caused by big movements, these present as abnormally large values and can be detected. When there is a sudden head turn or looking away, it is the arousal accompanied with the distraction that causes the movement and not the act of turning itself. However, it is always recommended to use videotapes to interpret these responses.

6.2. Specific challenges in eliciting HRV data during learning from children

1. In spite of the adjustable strap and getting the tightest fit, the PPG sensor for heart rate measurement sometimes did not achieve good contact with the wrist of the participant owing to their small wrist size, which resulted in data loss. In order to avert this, we used a small pad of cloth near the strap to enable better contact of the PPG sensor on wrist, especially for children of smaller build.
2. Optical Noise: the biggest challenge with HR measures from optical sensors is missing IBI data from motion noise. As a result, there were periods where the HR measures were not recorded because the child moved his hands a lot. While small movements do not affect this, sudden big movements result in a loss of data. Therefore, if the child is exerting mental effort but makes a movement during this period, there is always a risk of losing the data. Using video data and other contextual cues is important to decipher why data is missing or what may have ensued during this period.
3. In the event of noisy artifacts (since HR measures are more prone to movements than SCRs), it helps to look at the accelerometer data to see if there was a movement if video recordings do not present very obvious movements.

6.3. Other takeaways for implementing such triangulated frameworks

6.3.1. Task design

Having mixed difficulty tasks may give a good insight into whether the measures are truly responsive to randomly occurring difficulties/cognitive load and not just build up over time. Similarly, using a mix of long and short span tasks may best mimic learning in real-life situations.

6.3.2. Using the appropriate instrumentation

The choice of the wearable depends on the age group, task and the site of testing. Artifacts could be produced when electrodes are placed on fingers, and when an individual moves their knuckles [83]. Therefore, for Kindergarteners, we experienced that wearables with minimum instrumentation to collect as much forms of data and no distracting buttons or wires work best.

Furthermore, we used instruments that would be easier to employ in a real-classroom or play context.

6.3.3. Procedure

Having the experimenter spend a week with participants before the study, established rapport and removed any responses due to stranger anxiety.

It is recommended that a baseline be established before every task to estimate change in physiological measures.

6.3.4. Analysis

Normalizing the data to overcome inter-subject variability and analyse the difference between baseline measures and the task measures when comparing as a group is a key consideration. However, a true triangulation happens only when every subject's data is individually analysed and all measures are studied in relation to each other as shown in Figs. 10 and 11.

Selecting the section for analysis is also important. Usually, for GSR, one determines a minimum threshold and analyses responses above the threshold. If there are too many SCRs, an alternative way would be to look at the top 10% or top 20 SCRs [83]. However, analysing HRV may require segmenting the data and running a section-wise time and frequency domain analysis. Such an analysis would help detect the part of the task that contributed to cognitive stress if any.

6.4. Considerations and challenges in interpreting cognitive–affective states

Conducting such triangulated frameworks, especially with physiological data to understand cognitive–affective states come with additional considerations to bear in mind when making interpretations.

1. One of the biggest challenges has been mapping experience with expression/emotion. They are undoubtedly inextricably linked but there will be always be the challenge of mapping what kind of experiences map to what emotion in spite of other observable behaviours. Assuming that there exists such a mapping itself could be dangerous and misleading. Emotions are notoriously fuzzy, ill-defined, and possibly indeterminate [87]. Some may manifest outwardly and some may not. The ones that do manifest themselves are not necessarily similarly defined. This is due to individual differences and our own biases and prejudices in qualifying what someone maybe feeling.
2. Emotions are not always instantaneous [49] and therefore some of the coding of affect that characterizes a lot of the research may not truly capture the emotion/affect. Manual coding is often based on overt signs such as facial expressions, posture, eye gaze that while indicative may not always truly reflect the dynamics of the emotion.
3. Lack of a synchronized response or one-to-one relationship between physiological markers and cognitive–affective states The physiological indicators do not correspond directly or have a one-to-one with affective–cognitive states. As observed by Calvo et al. [49] no single *"sophisticated synchronized response that incorporates peripheral physiology, facial expression, speech, modulations of posture, affective speech, and instrumental action"* emerged for every affect.
4. This study was conducted with a small number of children in their schools making it hard to generalize the findings. Therefore, more studies in different contexts with more participants is needed.
5. We acknowledge that cognitive load itself is a fuzzy concept with different dimensions and definitions to it. Therefore, while we use cognitive load and mental effort synonymously in this paper, interpreting relationship of cognitive–affective states as a function of cognitive load must be done cautiously.
6. Finally, the task of defining learning itself has seen many a debate. Outlining the learning goals and then providing different ways to learn and demonstrate the learnt skills/concepts may help address this to some extent. Performance assessments should therefore provide for different ways to test for different dimensions of a concept. While, this framework is not endorsed to be a measurement of learning but rather of what ensues during learning, having such assessments of learning may provide useful insights to how/why/if certain states during learning may have contributed to a certain performance.

7. Future vision

We believe that identifying and triangulating sensitive measures, that supplement each other and provide insights into underlying cognitive–affective states, offers potential value in understanding and designing for learning. We foresee such a triangulated framework as a key step into creating smart classes and classrooms, through identification and assignment of cognitive–affective state terms to triangulated measures of physiology, behaviour and observation. We would like to call this "Triangulated Affective Learning" where a term is ascribed to the learner's state as s/he progresses through learning and this information is utilized by the teacher or even an intelligent tutor to match their responses and actions.

7.1. Comprehensive usability testing and building the right challenge in toys and learning interfaces

The framework could be expanded upon in the area of interaction design, where it could be applied to usability testing for learning interfaces and toys. Many behavioural responses such as yawns, sighs, turning away, frowns are more reliable indicators than the ratings/verbal feedback. While testing a construction toy or an interactive multimedia game, if the child displays an almost bored expression complemented by lack of SCRs and HRV and shows no difficulty in trying the product, it may suggest that it is within comfortable limits. Depending on what the product aims to accomplish, it is up to the designer to decide to make it more exciting, add more challenges and get the children to explore more. On the other hand, if this is a platform to learn

new content, then navigating the platform and getting used to it should be done with ease as indicated by low SCRs, low LF HRV and a general calm/neutral expression. Or, if we were to imagine a child tester displaying mild frustration with high SCRs and high LF HRV, the experimenter could more closely evaluate the point where this occurred and think of what aspect of the interaction/product feature may have brought about this response — did the toy have too many instructions or were there too many elements to remember (taxing working memory) or in the case of a website for children, did the child have a lot of distracting features like colours and cartoon typography (affecting inhibition and making it hard to focus) or did it require them to constantly shift between different features (taxing flexibility)? If data from different sources are triangulated, they may offer the experimenter a better understanding of the child's cognitive–affective state as they go through the usability testing. This coupled with what the child responded verbally, rated on a response scale may point to some pain points or good aspects of a design. This can aid the experimenter/designer to closely evaluate the causes and possible ways to rectify them and re-evaluate the new design.

7.2. Automated real-time feedback of cognitive–affective states in classrooms

We envision an automated and real-time measurement as well as feedback of learners' cognitive–affective states for each student in the classroom to teachers. If a child's SCRs are high in number and amplitude, with a higher LF component of HRV, but demonstrates a more curious or engrossed look, it may signal that the child is exploring or trying to understand a problem. Now if one were to contrast this with that of a child who exhibits the same SCRs and HRV values but has a sad expression or frustrated expression, then, it shows that the child is probably finding the content too challenging. At this point, as deemed appropriate by the teacher, the child may need some intervention in the form of feedback or a re-evaluation of the pedagogy on part of the teacher. Of course, once an intervention or remedial action is implemented, the same measures may offer an insight into whether this was effective at all.

Our study in the wild establishes that such an application of the framework is scalable. The apparatus and set-up is simple, with smart watches (that some students may already be wearing) and a video camera. In the future, we foresee that this could be done using machine learning classification and multi-modal sensing. Future research could explore establishment of new machine learning models from children of target age groups who differ developmentally.

7.3. Creation of affect-aware intelligent tutors

Our framework enables the design of intelligent affect-aware tutors that allow for adaptive and personalized learning. Such tutors will be endowed with the ability to identify fine differences between valence and arousal (for example, attributing frustration to negative valence, high arousal vs. boredom to negative valence low arousal). Further, we also envision that such affect-aware tutors will be able to match learners' learning styles/preferences with more effective personalized feedback. For example, a student who is outgoing, adventurous and enjoys hands-on learning may be assigned more project-based tasks and physical tinkering. While another student who enjoys a more instruction-based learning may be shown more demonstrations and written material. Video-recording children maybe a potential challenge in implementing such a framework. As such, we suggest some alternatives such as prior approval and consent from parents, grouping of children and positioning of cameras such that they only face children who have prior consent and record their images only. This data must be encrypted and stored in a way they are not shared with anyone other than the teachers. The information through recording should ideally information from the video tapes need to be coded to provide only the facial expressions. Alternatively, the video-recording can be only performed when the teacher and parents deem it to be useful in evaluating the learning process or using it to inform their own assessments and driving a change. While we do not envision that such tutors will dictate all of the learning in classrooms, we believe that they maybe a good way of delivering self-learning or learning online/distance mode. When used in classrooms, the data collected will definitely need to be channelled and run through teachers and other experts to ensure that the tasks assigned and reinforcements offered are in-line with the short and long term learning goals.

8. Conclusion

We explored the feasibility of obtaining direct measures of cognitive load using physiological sensors from Kindergarteners and then used observational data to make sense of the physiological measurements. We found potential GSR and HRV markers of cognitive load that are applicable to Kindergarteners. By triangulating these with observations, we were able to better explain how the child perceived the cognitive load. We believe that such an approach can be applied across age groups for learning and task performance. Given that there is a rapid proliferation of interactive educational and play applications, collecting a child's state during interaction with the application, could reveal insights into the application itself. In this direction, we explored learning behaviours and effect of interactivity as 11–13 year old children learnt concepts using different media. We found that while running a classroom study is affected by different confounds, having multiple sources of data helps understand some underlying phenomena and responses to learning through different media. We believe that such an understanding paves way for designing pedagogies, learning tools and other adaptive learning interfaces that are responsive to the learner's cognitive and affective states.

Acknowledgements

We would like to thank the participants, their parents and teachers for being so supportive throughout this study. We want to thank Niha Chakravarthy who helped with the video coding and transcription for the first study. We would also like to thank Shanaka Ransiri who helped with setting up the Empatica E4 and building a platform to collect data locally. And Vipula Dissanayake for helping with the content development for the classroom study. This research was supported partially by a doctoral fellowship from Singapore University of Technology and Design and from Ferrero Asia Ltd.

Appendix. Classroom study learning content

A.1. Build the first bridge

Learning goal: To understand how structures are designed, and more specifically the history and construction of the first bridge.

Experience: Students have to build the first bridge across the river (this bridge is no longer standing today). In the interactive version, they select materials and components to start building the bridge. Steps need to be selected in the correct sequence for the bridge to be successfully constructed, for example, if the bridge deck is chosen before piles are driven to support it

the deck splashes on the river and floats away. Audio narration provides guidance on what to consider if a step is selected in the wrong sequence *"try first adding some supports for the deck"* and also explains how materials were used to build the bridge after they have been selected *"wooden piles were driven 5 metres deep into the riverbed"*. In the passive version, they watch the bridge being built with audio narration providing the same information. Students in the Magic Window (MW) and IVR Condition have control over their view (similar to watching a 360 video) but were not able to otherwise interact with the environment or control the pace of the content.

A.2. The time travelling Mailman

Learning Goal: To understand what purposes structures are designed for and what makes a good structure.

Experience: Students get a first hand experience of what it was like to cross the river in the past. They are tasked with delivering a letter and parcel across the river at two timepoints in the 1800 s. In the first timepoint the bridge has not been built yet and they have to use a punt to cross the river. At the second timepoint students have to pay a toll and wait for the swing span of the bridge to open and close to let a boat pass before crossing the river, just like what foot passengers had to do at the time. Audio narration provides guidance on what actions to take and supplies information on the punt and bridge. In the interactive version, students are able to explore the environment at their own pace by selecting objects and teleporting to various locations using a button press or tapping the screen. In the passive version, students watched a first person view of events unfolding with audio narration providing the same information.

A.3. Learning questionnaires

Learning questionnaires were designed with teacher feedback, based on the specific learning goals for each content, as well as overall learning goals for the topic. The questions were:

1. Label the parts of the bridge and the materials they are made of.
2. List the steps involved in building the bridge.
3. How did people get across the river in (1862, 1866) and how was it different from today (2018)?
4. What do you need to think about when building a bridge?
5. Why do people build bridges?

References

[1] F.G. Paas, J.J. Van Merriënboer, Variability of worked examples and transfer of geometrical problem-solving skills: A cognitive-load approach, J. Educ. Psychol. 86 (1) (1994) 122.

[2] R.W. Picard, S. Papert, W. Bender, B. Blumberg, C. Breazeal, D. Cavallo, T. Machover, M. Resnick, D. Roy, C. Strohecker, Affective learning?a manifesto, BT Technol. J. 22 (4) (2004) 253–269.

[3] J.C. Read, S. MacFarlane, Using the fun toolkit and other survey methods to gather opinions in child computer interaction, in: Proceedings of the 2006 Conference on Interaction Design and Children, ACM, 2006, pp. 81–88.

[4] B.N. York, S. Loeb, One step at a time: The eects of an early literacy text messaging program for parents of preschoolers, Technical Report, National Bureau of Economic Research, 2014.

[5] W.H. Clune, P.A. White, Policy Effectiveness of Interim Assessments in Providence Public Schools. WCER Working Paper No. 2008-10., Wisconsin Center for Education Research (NJ1), 2008.

[6] P.A. Kirschner, Cognitive Load Theory: Implications of Cognitive Load Theory on the Design of Learning, 2002.

[7] T. De Jong, Cognitive load theory, educational research, and instructional design: some food for thought, Instr. Sci. 38 (2) (2010) 105–134.

[8] F.D. Pociask, G.R. Morrison, Controlling split attention and redundancy in physical therapy instruction, Educ. Technol. Res. Dev. 56 (4) (2008) 379–399.

[9] B.S. Hasler, B. Kersten, J. Sweller, Learner control, cognitive load and instructional animation, Appl. Cogn. Psychol. 21 (6) (2007) 713–729.

[10] B.F. Yuksel, K.B. Oleson, L. Harrison, E.M. Peck, D. Afergan, R. Chang, R.J. Jacob, Learn piano with bach: An adaptive learning interface that adjusts task difficulty based on brain state, in: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, ACM, 2016, pp. 5372–5384.

[11] J. Sweller, Cognitive load during problem solving: Effects on learning, Cogn. Sci. 12 (2) (1988) 257–285.

[12] R. Brunken, J.L. Plass, D. Leutner, Direct measurement of cognitive load in multimedia learning, Educ. Psychol. 38 (1) (2003) 53–61.

[13] F. Paas, J.E. Tuovinen, H. Tabbers, P.W. Van Gerven, Cognitive load measurement as a means to advance cognitive load theory, Educ. Psychol. 38 (1) (2003) 63–71.

[14] J. Leppink, A. van den Heuvel, The evolution of cognitive load theory and its application to medical education, Perspect. Med. Educ 4 (3) (2015) 119–127.

[15] K. Picho, A.R. Artino Jr, 7 deadly sins in educational research, 2016.

[16] R.E. Mayer, M. Hegarty, S. Mayer, J. Campbell, When static media promote active learning: Annotated illustrations versus narrated animations in multimedia instruction, J. Exp. Psychol. Appl. 11 (4) (2005) 256.

[17] A. Nakasone, H. Prendinger, M. Ishizuka, Emotion recognition from electromyography and skin conductance, in: Proc. of the 5th International Workshop on Biosignal Interpretation, 2005, pp. 219–222.

[18] C. Setz, B. Arnrich, J. Schumm, R. La Marca, G. Tröster, U. Ehlert, Discriminating stress from cognitive load using a wearable eda device, IEEE Trans. Inf. Technol. Biomed. 14 (2) (2010) 410–417.

[19] Y. Shi, N. Ruiz, R. Taib, E. Choi, F. Chen, Galvanic skin response (gsr) as an index of cognitive load, in: CHI'07 Extended Abstracts on Human Factors in Computing Systems, ACM, 2007, pp. 2651–2656.

[20] C.S. Ikehara, M.E. Crosby, Assessing cognitive load with physiological sensors, in: System Sciences, 2005. HICSS'05. Proceedings of the 38th Annual Hawaii International Conference on, IEEE, 2005, 295a–295a.

[21] J. Engström, E. Johansson, J. Östlund, Effects of visual and cognitive load in real and simulated motorway driving, Transp. Res. F 8 (2) (2005) 97–120.

[22] E. Ferreira, D. Ferreira, S. Kim, P. Siirtola, J. Röning, J.F. Forlizzi, A.K. Dey, Assessing real-time cognitive load based on psycho-physiological measures for younger and older adults, in: Computational Intelligence, Cognitive Algorithms, Mind, and Brain (CCMB), 2014 IEEE Symposium on, IEEE, 2014, pp. 39–48.

[23] E. Najafpour, N. Asl-Aminabadi, S. Nuroloyuni, Z. Jamali, S. Shirazi, Can galvanic skin conductance be used as an objective indicator of children?s anxiety in the dental setting? J. Clin. Exp. Dentistry 9 (3) (2017) e377.

[24] J.A. Posthumus, K. Böcker, M. Raaijmakers, H. Van Engeland, W. Matthys, Heart rate and skin conductance in four-year-old children with aggressive behavior, Biol. Psychol. 82 (2) (2009) 164–168.

[25] F. Bousefsaf, C. Maaoui, A. Pruski, Remote assessment of the heart rate variability to detect mental stress, in: Pervasive Computing Technologies for Healthcare (PervasiveHealth), 2013 7th International Conference on, IEEE, 2013, pp. 348–351.

[26] N. Hjortskov, D. Rissén, A.K. Blangsted, N. Fallentin, U. Lundberg, K. Søgaard, The effect of mental stress on heart rate variability and blood pressure during computer work, Eur. J. Appl. Physiol. 92 (1–2) (2004) 84–89.

[27] Z.B. Moses, L.J. Luecken, J.C. Eason, Measuring task-related changes in heart rate variability, in: Engineering in Medicine and Biology Society, 2007. EMBS 2007. 29th Annual International Conference of the IEEE, IEEE, 2007, pp. 644–647.

[28] D.J. McDuff, J. Hernandez, S. Gontarek, R.W. Picard, Cogcam: Contact-free measurement of cognitive stress during computer tasks with a digital camera, in: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, ACM, 2016, pp. 4000–4004.

[29] S.H. Fairclough, Fundamentals of physiological computing, Interact. Comput. 21 (1–2) (2008) 133–145.

[30] H.A. Simon, Motivational and emotional controls of cognition, Psychol. Rev. 74 (1) (1967) 29.

[31] D.A. Norman, Twelve issues for cognitive science, Cogn. Sci. 4 (1) (1980) 1–32.

[32] V. Vroom, Work and Motivation, Wiley New York, 1964.

[33] J.M. Keller, Motivational design of instruction, in: Instructional Design Theories and Models: An Overview of their Current Status, Vol. 1, 1983, pp. 383–434.

[34] J.M. Keller, Strategies for stimulating the motivation to learn, Perform. Improv. 26 (8) (1987) 1–7.

[35] P.L. Vail, Emotion: The on/off Switch for Learning, Modern Learning Press, 1994.

[36] C.S. Dweck, Motivational processes affecting learning, Amer. Psychol. 41 (10) (1986) 1040.

[37] C. Ames, J. Archer, Achievement goals in the classroom: Students' learning strategies and motivation processes, J. Educ. Psychol. 80 (3) (1988) 260.

[38] C.S. Dweck, E.L. Leggett, A social-cognitive approach to motivation and personality, Psychol. Rev. 95 (2) (1988) 256.
[39] M. Csikszentmihalyi, Finding flow, New York: Basic Books, 1997.
[40] B. Kort, R. Reilly, R.W. Picard, An affective model of interplay between emotions and learning: Reengineering educational pedagogy-building a learning companion, in: Advanced Learning Technologies, 2001. Proceedings. IEEE International Conference on, IEEE, 2001, pp. 43–46.
[41] G. Mandler, Mind and Body: Psychology of Emotion and Stress, WW Norton, 1984.
[42] M.R. Lepper, R.W. Chabay, Socializing the intelligent tutor: Bringing empathy to computer tutors, in: Learning issues for intelligent tutoring systems, Springer, 1988, pp. 242–257.
[43] Y. Matsubara, M. Nagamachi, Motivation system and human model for intelligent tutoring, in: Intelligent Tutoring Systems, Springer, 1996, pp. 139–147.
[44] A. De Vicente, H. Pain, Motivation diagnosis in intelligent tutoring systems, in: Intelligent Tutoring Systems, Springer, 1998, pp. 86–95.
[45] D. Whitelock, E. Scanlon, Motivation, Media and Motion: Reviewing a Computer Supported Collaborative Learning Experience, CITE REPORT, 1996.
[46] J.M. Keller, B. Keller, Motivational Delivery Checklist, Florida State University, 1989.
[47] A. Kapoor, S. Mota, R.W. Picard, et al., Towards a learning companion that recognizes affect, in: AAAI Fall symposium, 2001, pp. 2–4.
[48] R.A. Calvo, S.K. D'Mello, New Perspectives on Affect and Learning Technologies, Vol. 3, Springer Science & Business Media, 2011.
[49] R.A. Calvo, S. D'Mello, Affect detection: An interdisciplinary review of models, methods, and their applications, IEEE Trans. Affective Comput. 1 (1) (2010) 18–37.
[50] P. Ekman, Facial expression and emotion., Amer. Psychol. 48 (4) (1993) 384.
[51] S. DMello, A. Graesser, Dynamics of affective states during complex learning, Learn. Instr. 22 (2) (2012) 145–157.
[52] S.K. DMello, A. Graesser, Multimodal semi-automated affect detection from conversational cues, gross body language, and facial features, User Model. User-Adapt. Interact. 20 (2) (2010) 147–187.
[53] S. Craig, A. Graesser, J. Sullins, B. Gholson, Affect and learning: an exploratory look into the role of affect in learning with autotutor, J. Educ. Media 29 (3) (2004) 241–250.
[54] R. Sharma, V.I. Pavlović, T.S. Huang, Toward multimodal human–computer interface, in: Advances In Image Processing And Understanding: A Festschrift for Thomas S Huang, World Scientific, 2002, pp. 349–365.
[55] P. Blikstein, M. Worsley, Multimodal learning analytics and education data mining: using computational technologies to measure complex learning tasks, J. Learn. Anal. 3 (2) (2016) 220–238.
[56] L.J. Levine, N.L. Stein, Making sense out of emotion: The representation and use of goal-structured knowledge, in: Psychological and Biological Approaches to Emotion, Psychology Press, 2013, pp. 63–92.
[57] G.L. Clore, J.R. Huntsinger, How emotions inform judgment and regulate thought, Trends Cogn. Sci. 11 (9) (2007) 393–399.
[58] K. Fiedler, Affective states trigger processes, in: Theories of Mood and Cognition: A Users Guidebook, Lawrence Erlbaum Associates Publishers Mahwah, NJ, 2001, pp. 85–98.
[59] J.A. Russell, Core affect and the psychological construction of emotion, Psychol. Rev. 110 (1) (2003) 145.
[60] L.F. Barrett, Variety is the spice of life: A psychological construction approach to understanding variability in emotion, Cogn. Emotion 23 (7) (2009) 1284–1306.
[61] N.H. Frijda, Emotions, individual differences and time course: Reflections, Cogn. Emotion 23 (7) (2009) 1444–1461.
[62] C.E. Izard, Basic emotions, natural kinds, emotion schemas, and a new paradigm, Perspect. Psychol. Sci. 2 (3) (2007) 260–280.
[63] M.H. Immordino-Yang, A. Damasio, We feel, therefore we learn: The relevance of affective and social neuroscience to education, Mind Brain Educ. 1 (1) (2007) 3–10.
[64] C.S. Dweck, Messages that motivate: How praise molds students' beliefs, motivation, and performance (in surprising ways), in: Improving academic achievement, Elsevier, 2002, pp. 37–60.
[65] M. Csikszentmihalyi, I. Csikszentmihalyi, Beyond Boredom and Anxiety, Vol. 721, Jossey-Bass San Francisco, 1975.
[66] R. Pekrun, A.C. Frenzel, T. Goetz, R.P. Perry, The control-value theory of achievement emotions: An integrative approach to emotions in education, in: Emotion in Education, Elsevier, 2007, pp. 13–36.
[67] R.S. Baker, S.K. D'Mello, M.M.T. Rodrigo, A.C. Graesser, Better to be frustrated than bored: The incidence, persistence, and impact of learners cognitive–affective states during interactions with three different computer-based learning environments, Int. J. Hum.-Comput. Stud. 68 (4) (2010) 223–241.
[68] M.M.T. Rodrigo, R.S. d Baker, Comparing the incidence and persistence of learners affect during interactions with different educational software packages, in: New Perspectives on Affect and Learning Technologies, Springer, 2011, pp. 183–200.
[69] B. Woolf, W. Burleson, I. Arroyo, T. Dragon, D. Cooper, R. Picard, Affect-aware tutors: recognising and responding to student affect, Int. J. Learn. Technol. 4 (3–4) (2009) 129–164.
[70] R.W. Picard, S. Papert, W. Bender, B. Blumberg, C. Breazeal, D. Cavallo, T. Machover, M. Resnick, D. Roy, C. Strohecker, Affective learninga manifesto, BT Technol. J. 22 (4) (2004) 253–269.
[71] M. Scaife, Y. Rogers, Informing the design of a virtual environment to support learning in children, Int. J. Hum.-Comput. Stud. 55 (2) (2001) 115–143.
[72] A. Druin, The role of children in the design of new technology, Behav. Inf. Technol. 21 (1) (2002) 1–25.
[73] M. Garbarino, M. Lai, D. Bender, R.W. Picard, S. Tognetti, Empatica e3? a wearable wireless multi-sensor device for real-time computerized biofeedback and data acquisition, in: Wireless Mobile Communication and Healthcare (Mobihealth), 2014 EAI 4th International Conference on, IEEE, 2014, pp. 39–42.
[74] J.R. Simon, Reactions toward the source of stimulation, J. Exp. Psychol. 81 (1) (1969) 174.
[75] P.D. Zelazo, The dimensional change card sort (dccs): A method of assessing executive function in children, Nat. Protoc. 1 (1) (2006) 297.
[76] W. Schneider, A. Eschman, A. Zuccolotto, E-Prime: User's guide, Psychology Software Incorporated, 2002.
[77] P. Ekman, An argument for basic emotions, Cogn. Emotion 6 (3–4) (1992) 169–200.
[78] J.J. Braithwaite, D.G. Watson, R. Jones, M. Rowe, A guide for analysing electrodermal activity (eda) & skin conductance responses (scrs) for psychological experiments, Psychophysiology 49 (2013) 1017–1034.
[79] M. Benedek, C. Kaernbach, A continuous measure of phasic electrodermal activity, J. Neurosci. Methods 190 (1) (2010) 80–91.
[80] M.P. Tarvainen, J.-P. Niskanen, J.A. Lipponen, P.O. Ranta-Aho, P.A. Karjalainen, Kubios hrv–heart rate variability analysis software, Comput. Methods Programs Biomed. 113 (1) (2014) 210–220.
[81] S. Eldar, Y. Bar-Haim, Neural plasticity in response to attention training in anxiety, Psychol. Med. 40 (4) (2010) 667–677.
[82] J.T. Cacioppo, L.G. Tassinary, G. Berntson, Handbook of Psychophysiology, Cambridge University Press, 2007.
[83] E.B. Hedman, Thick Psychophysiology for Empathic Design (Ph.D. thesis), Massachusetts Institute of Technology, 2014.
[84] J.C. Read, Validating the Fun Toolkit: an instrument for measuring childrens opinions of technology, Cogn. Technol. Work 10 (2) (2008) 119–128, http://dx.doi.org/10.1007/s10111-007-0069-9, URL http://link.springer.com/article/10.1007/s10111-007-0069-9.
[85] D.C. Fowles, M.J. Christie, R. Edelberg, W.W. Grings, D.T. Lykken, P.H. Venables, Publication recommendations for electrodermal measurements, Psychophysiology 18 (3) (1981) 232–239.
[86] M.E. Dawson, A.M. Schell, C.G. Courtney, The skin conductance response, anticipation, and decision-making, J. Neurosci. Psychol. Econ. 4 (2) (2011) 111.
[87] R. Pekrun, L. Linnenbrink-Garcia, Academic emotions and student engagement, in: Handbook of Research on Student Engagement, Springer, 2012, pp. 259–282.